EveMRC: Two-Stage Bidirectional Evidence Modeling for Multi-Choice Machine Reading Comprehension

Hongshen Xu, *Graduate Student Member, IEEE*, Lu Chen, Liangtai Sun, Ruisheng Cao, Da Ma, and Kai Yu, *Senior Member, IEEE*

***Abstract*—Machine Reading Comprehension (MRC) requires a machine to answer questions after reading and comprehending the given documents. Multi-choice MRC is one of the most studied MRC tasks due to the convenience of evaluation and the diversity of question types. However, the interpretability of multi-choice MRC, especially in evidence extraction, remains underexplored because a correct answer may be derived by eliminating incorrect options rather than being supported by positive evidence. In this work, we propose a bidirectional evidence modeling framework, EveMRC, to enhance the explainability of Multi-choice MRC systems. Compared to previous works, our framework exclusively addresses the problem of bidirectional evidence selection, which not only selects positive evidence for the right answer but also selects negative evidence for wrong answers. The bidirectional evidence can also facilitate model decisions by incorporating it into a competition process. To avoid the high annotation cost of bidirectional evidence, our framework utilizes a novel weakly-supervised pipeline to train the evidence selector. Experimental results on four multi-choice MRC datasets demonstrate the effectiveness of our framework, which not only enhances the explainability of MRC systems but also improves their overall performance.**

***Index Terms*—Bidirectional evidence extraction, machine reading comprehension, the competition model.**

I. Introduction

Machine Reading Comprehension (MRC), which aims to teach machines to answer questions after reading given passages, is an important way to test the ability of intelligent systems to understand human language. With the help of many effective architectures [1], [2], [3] and pre-trained language models [4], [5], [6], reading comprehension systems are making rapid progress on many challenging datasets [7], [8], [9]. However, though current MRC systems could achieve better performances than humans, they lack explainability and are consequently prone to adversarial attacks [10], [11], [12].

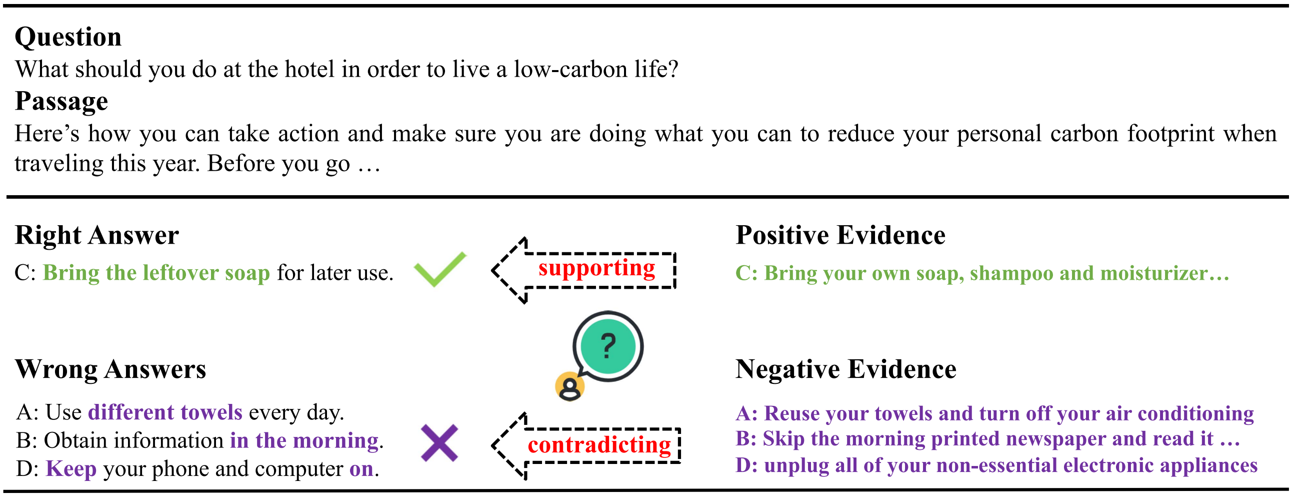
As the need to build more convincing MRC systems, research interest in explainability [13], [14] is rapidly growing. Models are required to expose the underlying mechanisms adopted to arrive at the final answers, whether by giving knowledge-based explanations, or by giving operational explanations such as the execution process of symbolic programs [15]. Many researchers focus on retrieving evidence sentences from passages as knowledge-based explanations due to their strong explainability even to non-specialists. While it is intuitive to extract surrounding sentences of answer span as evidence for extractive MRC tasks like SQuAD [7], it remains a challenge to extract evidence sentences for multi-choice MRC. On the one hand, the diversity of question types in multi-choice MRC results in more complex reasoning chains than other MRC tasks. Thus it requires the model's stronger reasoning abilities to select evidence. On the other hand, the flexibility of answer formats makes it more challenging in multi-choice MRC to retrieve evidence labels by manually annotating or heuristic rules. Consequently, there is a lack of multi-choice MRC datasets that provide annotated evidence for training. Utilizing unsupervised methods such as attention-based explanation [16] has gained more attention to provide evidence for multi-choice MRC systems.

Notably, most evidence selection methods for multi-choice MRC only focus on extracting evidence sentences for one predicted answer but ignore other answer candidates. These methods assume that the model's prediction is correct. However, when the model provides an incorrect answer, the corresponding erroneous evidence might further mislead users' judgment. We argue that providing evidence for all answer candidates will enhance both the explainability and reliability of MRC systems. There are a few works addressing the evidence selection for multiple answers by convincing MRC models to choose each answer candidate [17]. However, previous work ignores the crucial **evidence polarity** behind the multiple evidence selection problem. As shown in Fig. 1, not only the right answer but also the wrong answers are often semantically related to evidence. The right answer and the wrong answer may be distinguished, however, because only the right answer has positive evidence

Received 17 September 2023; revised 13 April 2024, 6 July 2024, and 22 January 2025; accepted 26 January 2025. Date of publication 14 February 2025; date of current version 28 February 2025. This work was supported in part by China NSFC Projects under Grant 62120106006, Grant U23B2057, and Grant 92370206, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102. The associate editor coordinating the review of this article and approving it for publication was Prof. Mark Hasegawa-Johnson. *(Corresponding authors: Lu Chen; Kai Yu.)*

The authors are with the X-LANCE Lab, Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xuhongshen@sjtu.edu.cn; chenlusz@sjtu.edu.cn; slt19990817@sjtu.edu.cn; 211314@sjtu.edu.cn; mada123@sjtu.edu.cn; kai.yu@sjtu.edu.cn).

Digital Object Identifier 10.1109/TASLPRO.2025.3542294



Question

What should you do at the hotel in order to live a low-carbon life?

Passage

Here's how you can take action and make sure you are doing what you can to reduce your personal carbon footprint when traveling this year. Before you go …

Right Answer

C: Bring the leftover soap for later use.

Positive Evidence

C: Bring your own soap, shampoo and moisturizer…

Wrong Answers

A: Use different towels every day.
B: Obtain information in the morning.
D: Keep your phone and computer on.

Negative Evidence

A: Reuse your towels and turn off your air conditioning
B: Skip the morning printed newspaper and read it …
D: unplug all of your non-essential electronic appliances

Fig. 1. An example from the RACE$^+$ dataset. Not only does the right answer have corresponding evidence, but also wrong answers may have evidence sentences. Both positive and negative evidence are crucial for improving the explainability and performance of MRC systems.

to support it and there exists negative evidence for each wrong answer to exclude them. Knowing the reason why answers are right or wrong provides a much stronger explainability for MRC systems. We also believe that this kind of **bidirectional evidence** can further enhance the MRC system to make more precise predictions.

To this end, we propose the setting of bidirectional evidence selection for multi-choice MRC systems. MRC systems are required to not only predict the right answer but also to generate evidence for all answer candidates in this setting. We further annotate an explainable multi-choice MRC benchmark, RACE$^-$, with bidirectional evidence for explainability evaluation. To address the aforementioned limitation of lack of evidence annotation, we propose EveMRC, a two-stage bidirectional evidence modeling framework for multi-choice reading comprehension inspired by the Competition Model (see Section II). To extract bidirectional evidence for both right and wrong answers, we utilize an erasure-based pseudo-evidence generating method to train our evidence selector. For each answer candidate, the evidence selector provides positive evidence to support it and negative evidence to contradict it. Furthermore, we enhance the MRC system with selected bidirectional evidence by incorporating the competition process among the evidence for different answers and evidence with different polarities.

Our main contributions are as follows:

- We propose a novel setting of bidirectional evidence selection, that captures the practical necessity of not only explaining why the answer is right but also explaining why the answer is wrong.
- We propose a two-stage evidence modeling framework for multi-choice MRC which not only models the bidirectional evidence but also models the competitive correlation among evidence.
- We conduct thorough experiments on four MRC datasets and the experimental results show that our framework not only provides stronger explainability but also improves the performance of MRC systems.

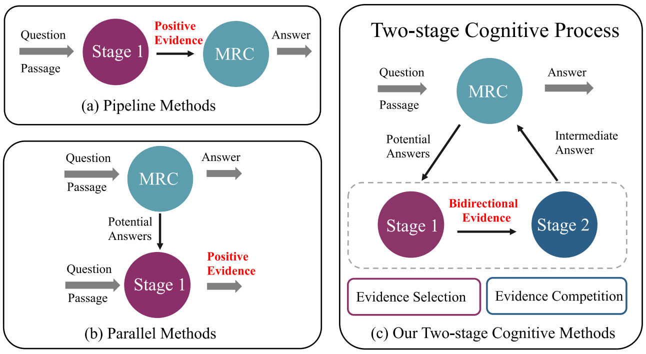


Fig. 2. Comparison between our proposed two-stage framework and previous works.

II. The Competition Model

The Competition Model is a psycholinguistic theory [18], [19], [20] which focuses on the competition process of sentence processing and language acquisition. It argues that humans understand a sentence by first searching for linguistic cues, such as word order, morphology, and semantic characteristics, to support each possible interpretation, eventually choosing the interpretation with the highest likelihood. Thus sentence processing can be viewed as a choice among different interpretations supported by different linguistic cues.

Inspired by the Competition Model, we argue that the human cognitive process of reading comprehension can also be modeled as a two-stage process: (i) Evidence Selection (ii) Evidence Competition. As shown in Fig. 2(c), we propose an explainable framework for machine reading comprehension, which builds a closed loop between MRC systems and two-stage evidence modeling. In stage 1, we collect evidence for each possible answer. In stage 2, we conduct evidence competition among competing answers and their supporting evidence. Stage 1 helps the MRC model to retrieve relevant information for answering the question, both positively and negatively. Stage 2 makes the answer judgement with supporting evidence, in a manner analogous to the human cognitive process of the Competition Model. We compare our framework with other explainable MRC methods. As shown in Fig. 2(a), pipeline methods [3], [21] first extract evidence from the passage, then substitute the passage with evidence for more efficient reading. Parallel methods [17] in Fig. 2(b) select evidence for each answer independently which brings strong explainability. However, both the above two types of methods ignore the evidence polarity as well as the exploitation of extracted evidence on model performance.

III. Related Work

A. Machine Reading Comprehension

Machine Reading Comprehension is a subtype of question-answering task that emphasizes the comprehension of the passage. Depending on the answer type, we can divide existing reading comprehension tasks into four categories: cloze style, multiple choice, span prediction, and free-form answer. Many large-scale MRC datasets [7], [22], [23], [24], [25], especially multi-choice and span-prediction styles, are proposed to evaluate the MRC systems. Moreover, with the blooming of powerful pre-trained language models [5], [6], [26], MRC systems could surpass human performance on several MRC datasets.

However, it remains a big challenge to expose the underlying mechanisms adopted to arrive at the final answers. Models are easily attacked by simple input perturbations [10], [12], [27] and thus are unable to meet the requirements for real-world applications, such as user trust, confidence and acceptance [15], [28]. Many works have been proposed to address the *explainability* problem of MRC systems. On the one hand, researchers build benchmarks with additional labeled explanation data for evaluating explainable MRC systems. On the other hand, many explanation methods have been applied to enhance the explainability of MRC systems. We will introduce more details about these works in the following two subsections.

B. Explainable MRC Datasets

Explainable MRC datasets are those benchmarks that provide additional explanation annotation of the answering reasoning process. Most explanations of those datasets are natural language sentences taken from original passages. While a few datasets include gold explanations that can be adopted as an additional training signal for explainable MRC models, more benchmarks only support the use of quantitative metrics for evaluating the explainability of MRC systems. For example, HotpotQA [9] provides sentence-level supporting facts for every question in both training and testing sets and introduces a leaderboard for evaluating the explanations. CoQA [25] contains free-form answers and each answer has a span-based rationale for training and testing. Instead, ExpMRC [29] annotated several datasets only for explainability evaluation.

C. Interpretation Methods

Due to the high annotation cost of evidence training labels, most research on interpreting MRC systems focuses on extracting evidence by unsupervised or weakly-supervised methods. One of the most typical unsupervised interpretation methods is attention-based explanation. There are many discussions about *whether attention is explainable*. Some researchers argue that attention does not necessarily correspond to the importance and may not be an optimal method to identify the attribution for an output [30], [31]. On the contrary, there are some researchers who hold positive attitudes [32], [33]. While there is no consensus about this topic, attention-based interpretation methods are still important baselines for many works. Apart from the attention mechanism, some works also seek to utilize other explainable modules for providing the explanation. For example, Moon et al. [34] proposes a Memory Graph Network to enable dynamic expansion of memory slots through graph traversals to improve the explainability of question answering systems. Cui et al. [35] proposes a Recursive Dynamic Gating mechanism that provides explanations through observing gating values. Unfortunately, it is nearly impossible to extract *bidirectional evidence* through the above self-explainable mechanisms. On the one hand, the absolute value of modular weights only represents the relevance rather than the polarity of the given evidence. On the other hand, it is impractical to associate evidence polarity with the numerical sign of modular weights due to its multi-aspect randomness such as model architecture.

Building upon evidence selection, many studies have further explored the use of selected evidence to support MRC systems. Min et al. [21] employs a sentence selector to extract evidence sentences, replacing the original passage for efficient reading comprehension. Wu et al. [36] masks out non-evidence sentences and utilizes attention scores to generate answers. Lin et al. [37] first trains a self-attention layer for evidence selection on top of the MRC system using supervision signals, further integrating the evidence attention with the original passage's self-attention to predict the final answer. However, these approaches predominantly rely on supervised data to train their explainers and overlook bidirectional evidence, focusing solely on selecting positive evidence to assist MRC systems. In contrast, we propose to adopt erasure-based methods [38], [39], [40] to derive attributions as pseudo labels for training our explainer. Furthermore, we investigate the polarity of output attributions to generate bidirectional evidence pseudo labels. By leveraging these pseudo labels to train our explainer, we achieve both performance improvement and stronger interpretability for MRC systems.

D. Explanation of Large Language Models

Large language models (LLMs) can achieve impressive performance on many new tasks by incorporating numerous in-context examples in their prompts [41]. Additionally, enhancing these examples with explanations can further improve the reasoning capabilities of LLMs across various tasks. For example, Chain-of-Thought (CoT) [42], [43] prompting provides intermediate reasoning steps as explanations in prompts, helping LLMs achieve state-of-the-art results in arithmetic, symbolic, and common-sense reasoning tasks. Besides, CoT methods have also been used to explain LLM answers to multiple-choice questions [44]. Lampinen et al. [45] further investigate the different impacts of pre- and post-answer explanations on model reasoning abilities. Auto-CoT [46] replaced human-annotated CoT prompting by automatically generating explanations and clustering demonstrations through LLMs. Moreover, prompting LLMs to generate explanations before producing answers [47], [48] can enable LLMs to reason with a two-stage explain-then-answer paradigm.

However, most aforementioned works on LLM explanations have limited relevance to smaller models. Some research [49], [50] focuses on using explanation data generated by LLMs to enhance the training of smaller models. Another related work [51] involves using post-hoc explanations from smaller models to construct in-context prompts for LLMs. We believe that *smaller models can also effectively assist LLMs in reasoning*. We experimented with using a small model to perform bidirectional evidence extraction as explanations during the reasoning process, which are then fed to the LLM for generating the final answer. Experimental results demonstrate that our framework can effectively leverage the evidence extracted by the small model to enhance the zero-shot reasoning capabilities of the LLM.

Recently it has been demonstrated that CoT reasoning may not be an accurate description of the reasons underlying an LLM's answer to a question [52]. The EveMRC model, by contrast, is explainable by design, and the explanations that it provides are therefore guaranteed to be relevant to the decision computed by the model. This is a desirable characteristic that may be difficult to achieve using LLMs.

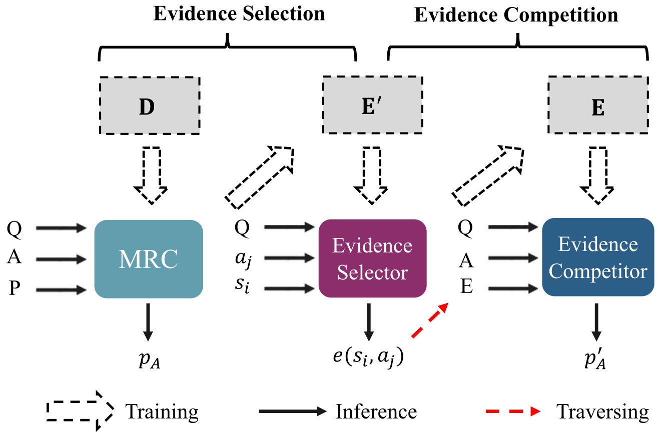


Fig. 3. Overview of our framework. We first train the MRC model with dataset D and generate pseudo-evidence label E′ to train the Evidence Selector. The evidence E generated by the Evidence Selector will be taken as inputs by the Evidence Competitor both in training and inference.

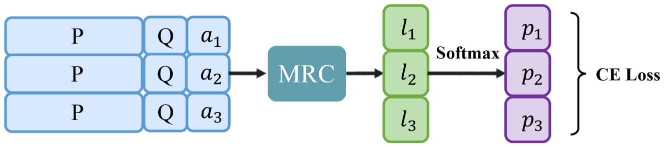


Fig. 4. Commonly used framework for multi-choice MRC. CE means Cross Entropy.

IV. Methods

A. Task Definition and System Overview

The task of multi-choice machine reading comprehension can be formalized as follows: given a reference passage $\mathrm{P} = \{s_1, s_2, \ldots, s_{l_p}\}$ composed of l_p sentences and a question Q, the model should select the right answer from the answer list $\mathrm{A} = \{a_1, a_2, \ldots, a_k\}$ with k answers (e.g., $k = 4$) which can be formalized as:

$$\hat{a} = \mathrm{argmax}_{a \in \mathrm{A}}\ P(a|\mathrm{P}, \mathrm{Q}).$$

Besides, most explainable MRC datasets only require the model to provide the evidence set composed of a few sentences to support the right answer. The right answer and the corresponding evidence set are denoted by a_r and $\mathrm{E_r}$, respectively.

We propose the setting of **bidirectional evidence selection** in the background of multi-choice MRC. For each answer a_j in the answer list A, the model is supposed to generate the corresponding evidence E_j. We define the evidence E_j which supports the answer a_j as **positive evidence**, the evidence E_j which contradicts the answer a_j as **negative evidence**. The evidence E_j can be either a few sentences or empty due to the fact that wrong answers do not always have negative evidence. Thus the model needs to provide the evidence set $\mathrm{E} = \{\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k\}$ for each question.

The overview of our framework is shown in Fig. 3. Due to the lack of evidence labels in most MRC datasets especially for bidirectional evidence, we explore the weakly supervised methods for evidence selection. More specifically, we first train the MRC model with the Multi-choice dataset D and generate pseudo-evidence label E′ with the probing from the trained MRC model (see Section IV-B). We will discuss more details about pseudo-evidence in Section IV-C. Furthermore, the trained Evidence Selector will give the evidential score $e(s_i, a_j)$ for the sentence-answer pair (s_i, a_j) and generate evidence E after traversing over all pairs (see Section IV-D). The evidence E will be taken as inputs by the Evidence Competitor both in training and inference. Finally, the answer probability from the MRC model and Evidence Competitor p_A, p'_A will be combined to get the final prediction in inference (see Section IV-E).

B. MRC Model With BCE Loss

The widely used framework for multi-choice MRC is shown in Fig. 4. Each answer in the answer list $\mathrm{A} = \{a_1, a_2, \ldots, a_k\}$ is concatenated with the passage P and question Q, resulting in k sequences. The model inputs k sequences separately, and outputs the logits $\mathrm{L} = \{l_1, l_2, \ldots, l_k\}$. Then the cross-entropy loss over the answer probability distribution $\mathrm{p} = \mathrm{softmax}(\mathrm{L})$ is adopted to train the model: $\mathcal{L}_{CE} = -log(\mathrm{p}_r)$, where r denotes the index of the right answer.

However, we believe that the above framework is not suitable as the MRC Model to generate the bidirectional pseudo-evidence label. We came to the conclusion based on the following two observations:

- The output probability $\mathrm{p_i}$ of answer a_i will be disturbed by other answers.

- The output logit l_i might be relevant to the interaction not only between the passage P and answer a_i but also between the passage P and question Q.

Concretely, assume that we mask one sentence s of the passage P to get the perturbed passage $\hat{\mathrm{P}}$, the logits and the probability distribution from the trained model will change from L, p to $\hat{\mathrm{L}}$, $\hat{\mathrm{p}}$, respectively. The value change is denoted by ΔL, Δp, respectively. If the sentence s is only related to answer a_1, the logit l_1 will decrease and the other logits remain unchanged. However, due to the softmax function, the probabilities for other answers will increase though only a_1 relates to the sentence.

Due to the fact that the logits are trained with CE loss after softmax, the logit l_i for each answer might contain the same bias eliminated by softmax. We simplify the logit l_i as follows:

$$\begin{aligned} l_i &= f(\mathrm{P}, \mathrm{Q}, a_i) \\ &\approx f_1(\mathrm{P}, \mathrm{Q}) + f_2(\mathrm{P}, a_i), \end{aligned} \tag{1}$$

where f, f_1, f_2 represents the interaction function among the inputs learned by neural networks. For the well-trained MRC models, the output logit l_i can be conceptualized as the learned interaction functions between the input passage P, the question Q, and the answer a_i. Additionally, it can be simplified to consider that the neural network learns the interaction function f_1 between the input passage P, the question Q, and the interaction function f_2 between variables the input passage P and the answer a_i, in order to generate the output logit l_i.

Then the logit change Δl_i can be formalized as:

$$\begin{aligned} \Delta l_i &= l_i(\hat{\mathrm{P}}, \mathrm{Q}, a_i) - l_i(\mathrm{P}, \mathrm{Q}, a_i) \\ &= f_1(\hat{\mathrm{P}}, \mathrm{Q}) - f_1(\mathrm{P}, \mathrm{Q}) \\ &\quad + f_2(\hat{\mathrm{P}}, a_i) - f_2(\mathrm{P}, a_i). \end{aligned} \tag{2}$$

Thus the logit change will suffer from the disturbance of interaction between passage P and question Q. When we mask out a sentence in the passage, we aim for the logit change to reflect the correlation between the masked sentence and the answer, rather than reflecting the correlation between the masked sentence and the question. This enables us to identify the evidence corresponding to each answer.

Instead, we train our model with binary cross-entropy loss to address the above two limitations. As shown in Fig. 5, the model inputs each sequence and judges each answer separately. Then the output logit l_i of answer a_i is passed to the sigmoid function: $\mathrm{p_i} = \mathrm{sigmoid}(l_i)$, and the final loss for each example is:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{k} [\mathbb{1}_{a_r}(a_i) log(\mathrm{p_i}) + \mathbb{1}_{a_r}(a_i) log(1 - \mathrm{p_i})],$$

where $\mathbb{1}_{a_r}(a_i) = 1$ if $a_i = a_r$ else 0, k denotes the number of answers, a_r denotes the right answer.

C. Generating Pseudo-Evidence Label

Over the past years, many approaches have been explored to interpret the MRC models where the attention-based methods [1], [53] are frequently used. However, it's almost impossible to determine the polarity of evidence generated by self-explainable mechanisms. In this work, we mainly focus on model-agnostic methods [54], [55]. The model-agnostic methods generate evidence solely based on model predictions and are applicable to any black-box models.

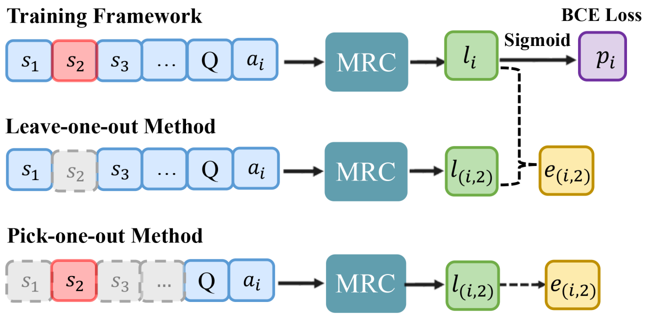


Fig. 5. The framework to train the MRC model and two model-agnostic methods to generate pseudo-evidence labels. $e_{(i,2)}$ represents the evidential score between the i-th answer and the sentence s_2. ($e_{(i,2)} = l_i - l_{(i,2)}$ in leave-one-out; $e_{(i,2)} = l_{(i,2)}$ in pick-one-out.).

We mainly introduce two model-agnostic methods in this paper. The first one is the erasure-based method which obtains an input subset's attribution by calculating the output change when erasing the subset. Following previous works [38], [40], we use the leave-one-out method to perform erasure and generate pseudo-evidence labels. The second one is a pickout-based method. By feeding input sentences into the MRC model separately [17], [56], we can calculate the probability that each sentence leads to or contradicts each answer.

As shown in Fig. 5, given the passage $\mathrm{P} = \{s_1, s_2, \ldots, s_{l_p}\}$ with l_p sentences, the model outputs the logit l_i for answer a_i when taking the full passage as inputs. For the leave-one-out method, when we erase the sentence s_2 and retain other sentences as inputs, the output logit is denoted by $l_{(i,2)}$. Thus, the attribution of sentence s_2 to answer a_i can be calculated by subtracting $l_{(i,2)}$ from l_i, i.e., $e_{(i,2)} = l_i - l_{(i,2)}$. For the pick-one-out method, the model only takes a single sentence s_2 as input and then outputs the logit, which can be viewed as the attribution between sentence s_2 and answer a_i directly.

After getting the attributions for all sentence-answer pairs, we sample the pseudo-evidence label for each example separately. The positive evidence from the right answer a_i can be sampled as:

$$(s'_+, a_+) = \underset{s_j}{\mathrm{argmax}} \in \mathrm{P}, a_i = a_r e_{(i,j)}.$$

Due to the fact that negative evidence leads to a decrease in answer probability, we sample one negative evidence from wrong answers as:

$$(s'_-, a_-) = \underset{s_j}{\mathrm{argmin}} \in \mathrm{P}, a_i \in \mathrm{A}^- e_{(i,j)},$$

where A^- denotes the set of wrong answers. Besides, we randomly sample one neutral sentence-answer pair (s'_n, a_n) as the neutral evidence.

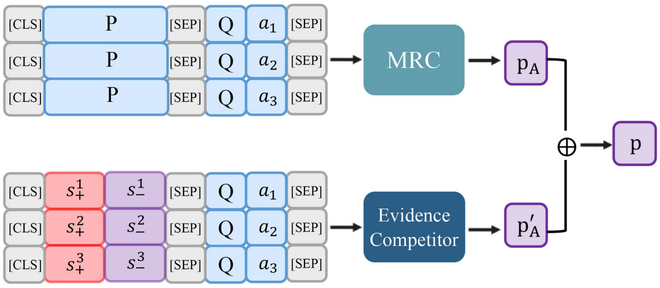


Fig. 6. The framework of Evidence Competition. The MRC model takes the full passage as input while the evidence competitor takes the bidirectional evidence instead. s_+^i, s_-^i represent the positive and negative evidence for answer a_i, respectively.

D. Evidence Selection

Given the pseudo-evidence label, we train the evidence selector to select evidence for each answer independently. We first construct the input sequence by concatenating sentence s_j, question Q and answer a_i. Then the input sequence is passed to a pre-trained encoder (e.g., BERT) to obtain the output embedding of the `[CLS]` token, which is finally passed to a linear layer for classification. We train two types of evidence selectors: unidirectional and bidirectional. The unidirectional selector adopts the positive sentence-answer pair (s'_+, a_+) and the neutral pair for training a binary classifier. The bidirectional selector adopts the positive, negative, and neutral pairs for training a three-class classifier. Both of the two selectors are trained by the pseudo-evidence label from the leave-one-out method.

E. Evidence Competition

The Competition Model suggests two types of evidence competition: (i) the competition between the positive and negative evidence of the answer, and (ii) the competition among evidence for all answers. To implement this, we search potential positive and negative evidence sentences for each answer. The output logits from the evidence selector which represent the positive, neutral, and negative evidential score of sentence s_j to answer a_i are denoted by $e_+(s_j, a_i)$, $e_n(s_j, a_i)$ and $e_-(s_j, a_i)$, respectively. Thus the positive and negative evidence of answer a_i can be obtained as follows:

$$s_+^i = \underset{s_j}{\text{argmax}} \in \mathrm{P} e_+(s_j, a_i),$$

$$s_-^i = \underset{s_j}{\text{argmax}} \in \mathrm{P} e_-(s_j, a_i).$$

After we obtain positive and negative evidence for all answers, the evidence competitor will use the evidence set rather than the full passage as input. As shown in Fig. 6, the framework of evidence competitor is similar to the MRC model. Instead, each answer a_i is concatenated with its positive evidence s_+^i and negative evidence s_-^i rather than the full passage. The evidence competitor is also required to predict the right answer a_r in training. For inference, the probability from the M RC model and the evidence competitor will be combined to get the final prediction, formally as:

$$\mathrm{p} = \alpha \mathrm{p}'_{\mathrm{A}} + (1 - \alpha)\mathrm{p}_{\mathrm{A}},$$

Where α indicates the combination ratio between the answer probabilities generated by the MRC model and the evidence competitor. A higher alpha value suggests a greater reliance on the judgment from the evidence competitor rather than the MRC model. Given that the evidence may encompass inherent noise, the utilization of alpha serves the purpose of considering the original output of the MRC model, thereby mitigating the negative influence of evidence noise.

V. Experiments

To evaluate the performance of our methods on bidirectional evidence selection and demonstrate the effectiveness of evidence competition, we conduct experiments on four multi-choice MRC datasets/benchmarks (see Section V-A). In Section V-C, we first evaluate the positive and negative evidence quantitatively, then report the improvement on answer accuracy by evidence competition. Finally, we will discuss the importance of negative evidence in Section V-D by presenting some cases.

A. Datasets

1) Multiple-Choice MRC: RACE [24]: RACE is collected from the English exams for middle and high school Chinese students, which consists of 27,933 passages and 97,687 questions.

DREAM [57]: DREAM is a dialogue-based dataset collected from English examinations, which contains 10,197 questions with 6,444 dialogues.

2) Explainable Multiple-Choice MRC: ExpMRC [29]: ExpMRC is an explainability evaluation benchmark. We use the multi-choice subsets: **RACE**$^+$ and **C**^{3} for evaluation and corresponding datasets: RACE and C^3 [58] for training. RACE$^+$ and C^3 contain 1,125 English questions and 1,005 Chinese questions with positive evidence, respectively.

3) RACE$^-$*:* To evaluate negative evidence, we propose RACE$^-$, a RACE-style testbed annotated with bidirectional evidence. To reduce the annotation cost, we choose the RACE$^+$ for re-annotation due to the fact that it already filters the invalid questions and annotates positive evidence. Based on the publicly available development set of RACE$^+$, the authors of this paper annotated the negative evidence. Specifically, given the passage, question, and answer list with the golden answer, we try to copy-and-paste a few sentences as the negative evidence for each wrong answer. The negative evidence is required to be selected as a sufficient condition to exclude the wrong answer, which could also be empty. The dataset statistics are shown in Table II.

B. Experimental Details

To evaluate our methods, we use two pre-trained language models: BERT$_{base}$ [5] and ALBERT$_{xxlarge\_v2}$ [5] of which the implementation is based on the Transformers.[1] The MRC

[1] https://github.com/huggingface/transformers

TABLE I
Experimental Results on RACE$^+$ Dataset

Model	**RACE$^+$ (dev)**			**RACE$^+$ (test)**			**C^3 (dev)**			**C^3 (test)**		
	Ans.	**Evi.**	**All**	**Ans.**	**Evi.**	**All**	**Ans.**	**Evi.**	**All**	**Ans.**	**Evi.**	**All**
PLM base-level Baselines♣												
Most Similar Sent.	62.4	36.6	28.2	59.8	34.4	26.3	68.7	57.7	47.7	66.8	52.2	41.2
Most Similar Sent. w/Ques.	62.4	44.5	31.5	59.8	41.8	27.3	68.7	62.3	47.3	66.8	57.4	42.3
Pseudo-data training	63.6	45.7	31.7	60.1	43.5	27.1	70.9	59.9	43.5	69.0	57.5	40.6
PLM large-level Baselines♣												
Most Similar Sent.	69.0	37.6	29.9	68.1	36.8	28.9	73.1	59.4	49.9	72.0	52.7	43.9
Most Similar Sent. w/Ques.	69.0	48.0	36.8	68.1	42.5	31.3	73.1	63.2	50.9	72.0	58.4	46.0
Pseudo-data training	**69.0**	45.9	32.6	**70.4**	41.3	30.8	**76.4**	64.3	50.7	**74.4**	59.9	47.3
Our Method												
PLM-base + EveMRC	66.7	**58.5**	**47.2**	66.7	**52.5**	**40.7**	71.3	**69.3**	**57.7**	67.6	**65.3**	**51.9**

Ans.: answer accuracy. **Evi.**: F1 score between golden evidence label and selected evidence sentences. All reflects the correctness of both answer and its evidence. PLM: pre-trained language model, e.g., BERT. ♣: results are taken from [29].

TABLE II
Dataset Statistics of RACE$^-$

Dataset	**Right Answers**	**Wrong Answers**			**Total**
		Has Evi.	**No Evi.**	**All**	
RACE$^-$	561	1185	472	1657	2218

The number of right answers is equal to the number of questions.

TABLE III
Hyperparameters of Training

Dataset		RACE		DREAM		C^3
Model		BERT	ALBERT	BERT	ALBERT	BERT
max seq-length	MRC	512	512	512	512	512
	ES/EC	200	200	200	200	200
learning rate	MRC/EC	3e-5	1e-5	2e-5	2e-5	3e-5
	ES	3e-5	3e-5	3e-5	3e-5	3e-5
batch size	MRC/EC	32	8	24	8	32
	ES	32	32	32	32	32
epoch	MRC/EC	3	3	8	2	5
	ES	2	2	2	2	2
warmup ratio	All	0.1	0.1	0.1	0.1	0.1
optimal alpha	-	0.5	0.5	0.2	0.3	0.4

ES: evidence selector. EC: evidence competitor.

model, evidence selector, and evidence competitor all employ a pre-trained language model as the encoder, and a single-layer linear network as the output layer. For each example in all datasets, we sample one pseudo-evidence for negative, positive, and neutral examples separately to train the evidence selector. The evidence competitor takes the top 3 and 1 sentence as evidence for RACE and DREAM, respectively. We search for the best coefficient α between 0.1 and 0.5 on the development set. We use AdamW [59] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and without weight decay. The training hyperparameters are shown in Table III. We search for the best weighting coefficient of probability combination on the dev set which ranges from 0.1 to 0.5 with 0.1 as the interval. For the MRC model with BCE loss, we simply prepare the features as normal MRC and compute loss for each answer separately with BCE, which means we adopt the same hyperparameters to train the BCE and CE MRC models. We follow the experimental settings from the leaderboards and corresponding papers. If there is no relevant information, we train the model three times and pick the model with the best accuracy on the dev set. We use 4 NVIDIA 2080Ti for the experiments with BERT$_{base}$ and 4 NVIDIA A10 for ALBERT$_{xxlarge}$.

C. Main Results

1) *Positive Evidence Selection:* To evaluate the positive evidence, we first submit our system to the leaderboard of ExpMRC which achieves the best results on two multi-choice MRC test sets. The submitted system consists of an MRC model, a unidirectional evidence selector, and an evidence competitor, which simultaneously selects evidence and answers the question. ExpMRC uses F1 score and accuracy to evaluate the evidence and answer, respectively. As shown in Table I, our system achieves significant improvement over the same base-level baselines on evidence selection (+10.9 F1, +7.4 F1 for RACE$^+$ and C^3 respectively). Although we did not submit the results based on the large-level models, our methods with PLM-base also improve the performance of evidence selection substantially over large-level baselines (+10.3 F1, +5.2 F1 for RACE$^+$ and C^3 respectively).

Considering that the above results of evidence selection will be influenced by answer accuracy, we further select evidence for the golden answer and report the results on RACE$^+$. Moreover, due to the fact that evidence in RACE$^+$ is a set of sentences, we adopt precision@k and recall@k as the additional metrics, which represent the precision and recall of the generated evidence labels, respectively, when k sentences are predicted as evidence.

As shown in Table IV, our unidirectional evidence selector achieves the best results on all metrics. The bidirectional selector performs worse than the unidirectional one which is probably caused by the introduction of noise by negative pseudo-evidence labels. Nevertheless, our two selectors both surpass the baselines by a large margin on the recall scores. It can be inferred that the sentence-level metrics provide a different view to evaluate the evidence.

2) *Negative Evidence Selection:* Due to the fact that negative evidence of wrong answers might not exist, the problem of negative evidence selection is analogous to question answering with

TABLE IV
POSITIVE EVIDENCE EVALUATION WITH GOLDEN OPTION ON THE DEVELOPMENT SET OF RACE$^+$

Methods	RACE$^+$ (dev)				
	Evi. F1	**P@1**	**R@1**	**R@3**	**R@5**
BERT$_{base}$					
Pick-one-out	50.8	53.1	43.8	62.8	71.7
Leave-one-out	55.5	58.5	48.8	61.2	64.0
Bidirectional ES	57.9	62.4	51.6	69.8	78.2
Unidirectional ES	**63.1**	**69.2**	**57.5**	**77.7**	**86.1**
ALBERT$_{xxlarge}$					
Pick-one-out	44.6	44.9	37.2	54.7	64.6
Leave-one-out	59.2	65.4	52.5	67.9	71.4
Bidirectional ES	66.0	71.5	59.5	79.2	84.8
Unidirectional ES	**69.4**	**76.5**	**63.4**	**83.7**	**90.1**

P@k / R@k represent precision / recall of the sentence-level evidence labels, respectively for top k predicted evidence sentences. ES: evidence selector.

TABLE V
BIDIRECTIONAL EVIDENCE EVALUATION ON RACE$^-$

Methods	**Pos. Evi. F1**	**Neg. Evi. F1**	**All Evi. F1**
BERT$_{base}$			
Pick-one-out	46.8	28.5	33.2
Leave-one-out	51.7	35.2	39.4
Bidirectional ES	52.4	**38.5**	**41.3**
Unidirectional ES	**57.2**	35.9	41.2
ALBERT$_{xxlarge}$			
Pick-one-out	44.0	28.6	32.5
Leave-one-out	59.1	46.4	49.6
Bidirectional ES	64.6	**49.2**	**53.1**
Unidirectional ES	**68.1**	35.2	43.5

TABLE VI
COMPARISON BETWEEN ATTENTION-BASED METHODS AND OUR EVIDENCE SELECTOR

Methods	**Pos. Evi. F1**	**Neg. Evi. F1**	**All Evi. F1**
Attention p2p	14.8	28.5	24.7
Attention a2p	15.5	28.6	25.1
Attention q2p	45.8	31.9	36.0
Attention p2q	52.2	32.2	37.7
Bidirectional ES	52.4	**38.5**	**41.3**
Unidirectional ES	**57.2**	35.9	41.2

P, Q, a represents passage, question, and answer, respectively. P2q represents the attention scores which take passage words as keys and question works as queries.

TABLE VII
EXPERIMENTAL RESULTS ON RACE AND DREAM DATASETS

Model	**RACE**		**DREAM**	
	dev Acc	**test Acc**	**dev Acc**	**test Acc**
BERT$_{base}$	66.5	65.2	63.4	63.2
+Evidence Competition	68.1	67.2	64.5	63.8
ALBERT$_{xxlarge}$	87.5	86.4	89.2	88.5
+Evidence Competition	88.4	87.6	90.0	89.4

unanswerable questions [8]. For each wrong answer, models are required to give negative evidence and abstain from explaining when it is unavailable. We use the evidential score for all methods to judge whether to abstain. Concretely, We treat the problem of negative evidence selection as question answering with unanswerable questions such as SQuAD2.0. For the pick-one-out and leave-one-out methods, we use the evidential score between the sentences and answers as the explain-or-not probability. For the evidence selector, we use the output probability of selected evidence to make the classification. We tune the best classification threshold separately for each model to maximize the F1 score as SQuAD2.0.

Table V shows the evaluation result of bidirectional evidence on RACE$^-$. We evaluate the positive and negative evidence on right and wrong answers, respectively. The overall results are averaged over all answers to all questions equally. For both positive and negative evidence, we use the F1 score as the evaluation metric which is the same as RACE$^+$. To ensure the fairness of results, we use the same MRC model for all methods to predict the answer. The results show that our bidirectional evidence selector achieves the best result on the F1 score of negative evidence. The unidirectional evidence selector performs best on positive evidence but the bidirectional one performs best on overall results. Furthermore, we observed that when we changed the base of the unidirectional evidence selector from BERT$_{base}$ to ALBERT$_{xxlarge}$, its ability to select negative evidence did not improve. This is because the unidirectional evidence selector is trained using positive pseudo labels, which only allows it to find some negative evidence through semantic relevance. It does not genuinely learn how to extract negative evidence, thus there exists an upper limit to its negative evidence score. Switching to a stronger model does not increase this upper limit.

3) Comparison With Attention-Based Methods: Attention mechanisms have been frequently used for revealing the prediction process with attended sentences [1]. Thus we adopt the attention-based methods for comparison. Concretely, we use different parts of the attention weight matrix as the scoring method to select evidence sentences. In Table VI, we represent the passage with p, the concatenation of question and answer with q, and the concatenation of all the three with a. We use the cross-attention weights matrix $W \in \mathbb{R}^{(l_p+l_q)\times(l_p+l_q)}$ in the last transformer layer of BERT$_{base}$ to generate evidence. More specifically, for the p2q method, we pool the attention scores with the i-th passage word p_i as the key and all the question and answer words as queries to score each passage word. After we get the importance score for each passage word, we achieve the sentence score by pooling the words' score within the sentence. The other methods are similar to the p2q method. As shown in Table VI, our bidirectional evidence selector achieves significant improvement over all the attention methods on negative evidence selection. Although the attention-based methods produce competitive results on positive evidence selection, our unidirectional selector also outperforms it by a large margin.

4) Evidence Competition: We evaluate the effect of evidence competition on two datasets: RACE and DREAM. Table VII shows the overall results. The baseline results are taken from

TABLE VIII
ABLATION STUDY OF EVIDENCE COMPETITION ON RACE

Methods	RACE	
	Dev Acc	Test Acc
BERT$_{base}$	66.5	65.2
+ Unidirecitonal EC	68.3 (+1.8)	66.9 (+1.7)
+ Negative EC	67.4 (+0.9)	65.8 (+0.6)
+ Positive EC	67.9 (+1.4)	66.6 (+1.4)
+ Bidirectional EC	68.1 (+1.6)	67.2 (+2.0)
+ Bidirectional EC†	**68.9 (+2.4)**	**67.5 (+2.3)**

† denotes evidence competition with bidirectional evidence from two separate evidence selectors.

the corresponding leaderboards and papers [40]. All of our methods are significantly better than the corresponding baselines with a p-value < 0.05 (t-test). As shown in Table VII, our methods improve the model performance by +2.0% and +1.1% for BERT$_{base}$ and ALBERT$_{xxlarge}$, which demonstrates that our method can help both a trivial baseline as well as a competitive baseline. However, the improvement on RACE is higher than on DREAM. The reason may be that the average sentence number of each passage in DREAM is only half of that in RACE so it will be less helpful for answering the question in DREAM to locate evidence.

To determine the contribution of positive and negative evidence, we train the evidence competitor that only takes the positive or negative evidence as inputs for the ablation study. As shown in Table VIII, the positive evidence gives the main contribution. It is reasonable due to the fact that only negative evidence is insufficient to answer the question in most cases. The bidirectional evidence competition achieves the best results by further competing between positive and negative evidence. We also incorporate the results of unidirectional evidence competition which takes the positive evidence from the unidirectional evidence selector for comparison. Unidirectional evidence competition achieves similar results to bidirectional evidence competition with the help of more accurate positive evidence (slightly better on the dev set and slightly worse on the test set). We attribute this to the bidirectional evidence selector needing to learn to extract both positive and negative evidence. This dual learning task leads to a decline in the quality of positive evidence extraction, offsetting the performance gains brought by negative evidence. To address this, We trained a positive evidence selector and a negative evidence selector using positive pseudo labels and negative pseudo labels, respectively, for extracting positive and negative evidence. We then use the bidirectional evidence from two separate evidence selectors for evidence competition. As shown in the last row of Table VIII, this approach can result in greater performance improvements, which further demonstrate the effectiveness of bidirectional evidence competition.

5) Evidence Competition for LLMs: We further attempted to enhance the zero-shot reasoning capabilities of large language models (LLMs) using evidence competition. We chose ALBERT as the bidirectional evidence selector to extract bidirectional evidence. This evidence was then used to assist LLMs (Llama 2-chat 7b [60] and Llama 2-chat 13b) in reasoning. This process is analogous to how humans skim and then closely read an article. The small model performs an initial read-through of the document to extract evidence, which is then provided to the LLMs for a detailed read-through and final answer generation. As shown in Table IX, our framework significantly enhances the zero-shot reasoning performance on RACE for both Llama 2-chat 7b and Llama 2-chat 13b. Furthermore, providing bidirectional evidence yields better performance improvements compared to providing unidirectional evidence. All prompts used are listed in the Appendix. The sampling temperature is set to 0 for all experiments.

TABLE IX
EXPERIMENTAL RESULTS OF EVIDENCE COMPETITION FOR LLMS ON RACE

Methods	RACE	
	Dev Acc	Test Acc
Llama 2-chat 7b	53.8	54.2
+ Unidirecitonal EC	56.7 (+2.9)	57.4 (+3.2)
+ Bidirectional EC	57.3 (+3.5)	58.1 (+3.9)
Llama 2-chat 13b	58.2	58.7
+ Unidirecitonal EC	61.3 (+3.1)	61.5 (+2.8)
+ Bidirectional EC	62.1 (+3.9)	62.4 (+3.7)

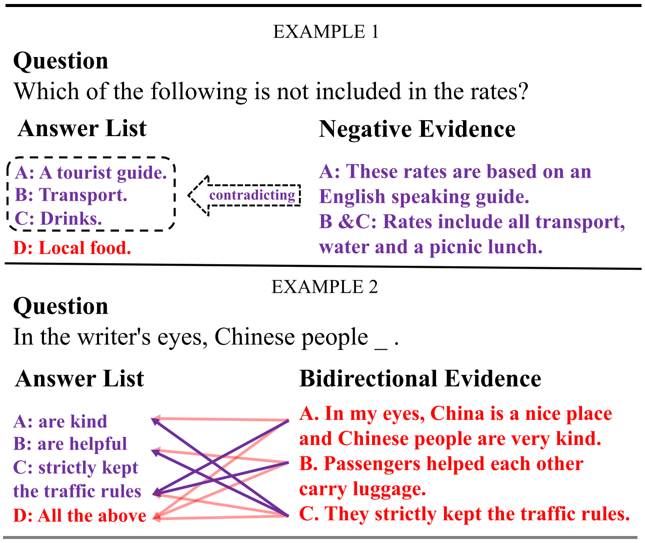

EXAMPLE 1
Question
Which of the following is not included in the rates?
Answer List
A: A tourist guide.
B: Transport.
C: Drinks.
D: Local food.
contradicting
Negative Evidence
A: These rates are based on an English speaking guide.
B &C: Rates include all transport, water and a picnic lunch.
EXAMPLE 2
Question
In the writer's eyes, Chinese people _ .
Answer List
A: are kind
B: are helpful
C: strictly kept the traffic rules
D: All the above
Bidirectional Evidence
A. In my eyes, China is a nice place and Chinese people are very kind.
B. Passengers helped each other carry luggage.
C. They strictly kept the traffic rules.

Fig. 7. Two examples with bidirectional evidence in BERT-RACE. The right answer and positive evidence are in red, and the wrong answers and negative evidence are in purple.

D. Case Study

To further investigate the importance of negative evidence and the effect of evidence competition, we focus on those examples in which the BERT$_{base}$ model failed to give the right answer but corrected its prediction with evidence competition.

As shown in example 1 of Fig. 7, the right answer "Local food" does not have any positive evidence but each wrong answer has its corresponding negative evidence. The BERT$_{base}$ model chose the wrong answer "A tourist guide", possibly due to the matching similarity between the "A tourist guide" with "English speaking `guide"`. However, our evidence selector successfully determined the polarity of all negative evidence for each wrong answer. Then the evidence competitor uses the negative evidence to exclude all wrong answers and finally arrive at the right answer.

For example 2 in Fig. 7, both the right answer and wrong answer have positive evidence to support them. It will be confusing for the model to choose the best answer from the answers each of which has its own positive evidence. By considering negative evidence, our method gives the best prediction `"All the above"`. Even if the wrong answer has positive evidence, it has negative evidence too. The answer with the most positive evidence and the least negative evidence would be the best choice.

We came to the conclusion that *negative evidence contributes to answering the question by excluding wrong answers*, especially in those cases where it is hard to retrieve positive evidence for the right answer or it is hard to distinguish plenty of positive evidence for all candidate answers.

VI. Conclusion

With the emerging research interest in explainable MRC systems, this paper proposes an explainable MRC framework for evidence extraction and evidence competition. We propose the setting of bidirectional evidence selection and tackle the problem of lacking labeled evidence data by applying weakly supervised methods. The experimental results show the effectiveness as well as the strong explainability of our framework. In the future, we will explore more unsupervised methods to utilize and enhance the explainability of MRC systems.

Appendix

We use the prompt below for the multi-choice MRC task.

```
TASK DESCRIPTION:
You will be given a passage to read.
After reading the passage, you need to
answer a corresponding question with
four answer options. Determine the
correct answer by choosing A, B, C, or
D, and provide the corresponding letter.

  TASK INPUT:
  Article:
  {article}

  Question:
  {question}

  Options:
  {option_text}

  TASK OUTPUT:
  Answer:
  A or B or C or D
```

We use the prompts below for unidirectional and bidirectional evidence competition, respectively.

```
TASK DESCRIPTION:
You will be given a passage to read.
After reading the passage, you need to
answer a corresponding question with
four answer options. The possible evi-
dence of options is also provided. Evi-
dence consists of sentences that support
the option as the correct answer. Note
that some evidence may be irrelevant
to the options, so you should rely on
the passage itself rather than solely on
the evidence to make your final choice.
Finally, based on the passage and evi-
dence, determine the correct answer by
choosing A, B, C, or D, and provide the
corresponding letter.

  TASK INPUT:
  Article:
  {article}

  Question:
  {question}

  Options:
  {option_text}

  Possible Evidence:
  {positive_evidence}

  TASK OUTPUT:
  Answer:
  A or B or C or D
```

```
TASK DESCRIPTION:
You will be given a passage to read.
After reading the passage, you need
to answer a corresponding question
with four answer options. The possible
evidence of options is also provided.
Positive evidence consists of sentences
that support the option as the correct
answer, while negative evidence con-
sists of sentences from the passage
that refute the option as the correct
answer. Note that some evidence may be
irrelevant to the options, so you should
rely on the passage itself rather than
solely on the evidence to make your
final choice. Finally, based on the
passage and evidence, determine the
correct answer by choosing A, B, C, or
D, and provide the corresponding letter.
```

```
TASK INPUT:
Article:
{article}

Question:
{question}

Options:
{option_text}

Possible Positive Evidence:
{positive_evidence}

Possible Negative Evidence:
{negative_evidence}

TASK OUTPUT:
Answer:
A or B or C or D
```

References

[1] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bi-directional attention flow for machine comprehension," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[2] A. W. Yu et al., "QANet: Combining local convolution with global self-attention for reading comprehension," in *Proc. 6th Int. Conf. Learn. Representations*, 2018

[3] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, and X. Zhou, "DCMN: Dual co-matching network for multi-choice reading comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9563–9570.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc 2019 Conf. North Amer. Chap. Assoc. Comput. Linguist.: Hum. Lang. Technol.*, vol. 1, 2019, pp. 4171–4186.

[5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Representations*, 2020.

[6] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *8th Int. Conf. Learn. International Representations*, 2020.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000 questions for machine comprehension of text," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.

[8] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, 2018, vol. 2, pp. 784–789.

[9] Z. Yang et al., "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proc. 2018 Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2369–2380.

[10] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. 2017 Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2021–2031.

[11] Y. Wang and M. Bansal, "Robust machine comprehension models via adversarial training," in *Proc. 2018 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Jun. 2018, pp. 575–581. [Online]. Available: https://aclanthology.org/N18-2091

[12] C. Si, Z. Yang, Y. Cui, W. Ma, T. Liu, and S. Wang, "Benchmarking robustness of machine reading comprehension models," in *Proc. Findings Assoc. Comput. Linguistics*, Aug. 2021, pp. 634–644. [Online]. Available: https://aclanthology.org/2021.findings-acl.56

[13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.

[14] B. Kratzwald, S. Feuerriegel, and H. Sun, "Learning a cost-effective annotation policy for question answering," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3051–3062.

[15] M. Thayaparan, M. Valentino, and A. Freitas, "A survey on explainability in machine reading comprehension," 2020, *arXiv:2010.00389*.

[16] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, "Attention interpretability across NLP tasks," 2019, *arXiv:1909.11218*.

[17] E. Perez, S. Karamcheti, R. Fergus, J. Weston, D. Kiela, and K. Cho, "Finding generalizable evidence by learning to convince Q&A models," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, pp. 2402–2411.

[18] E. Bates and B. MacWhinney, "Functionalist approaches to grammar," *Lang. Acquisit.: State Art*, pp. 173–218, 1982.

[19] E. Bates and B. MacWhinney et al., "Functionalism and the competition-model," *Crosslinguist. Study Sentence Process.*, vol. 3, pp. 73–112, 1989.

[20] B. MacWhinney, "Second language acquisition and the competition model," in *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers, 1997, pp. 113–142.

[21] S. Min, V. Zhong, R. Socher, and C. Xiong, "Efficient and robust question answering from minimal context over documents," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2018, pp. 1725–1735. [Online]. Available: https://aclanthology.org/P18-1160

[22] M. Richardson, C. J. Burges, and E. Renshaw, "MCTEST: A challenge dataset for the open-domain machine comprehension of text," in *Proc. 2013 Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 193–203.

[23] K. M. Hermann et al., "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.

[24] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-scale reading comprehension dataset from examinations," in *Proc. 2017 Conf. Empirical Methods Natural Lang. Process.*, pp. 785–794, 2017.

[25] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 249–266, 2019.

[26] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment in," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8018–8025.

[27] S. Goyal, S. Doddapaneni, M. M. Khapra, and B. Ravindran, "A survey of adversarial defenses and robustness in NLP," *ACM Comput. Surv.*, vol. 55, no. 14s, pp. 1–39, 2023.

[28] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *Proc. Int. Joint Conf. Artif. Intell. Workshop Explainable AI*, 2017, pp. 8–13.

[29] Y. Cui, T. Liu, W. Che, Z. Chen, and S. Wang, "ExpMRC: Explainability evaluation for machine reading comprehension," *Heliyon*, vol. 8, no. 4, 2022.

[30] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Jun. 2019, pp. 3543–3556. [Online]. Available: https://aclanthology.org/N19-1357

[31] S. Serrano and N. A. Smith, "Is attention interpretable?," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 2931–2951. [Online]. Available: https://aclanthology.org/P19-1282

[32] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, pp. 11–20, 2019.

[33] J. Bastings and K. Filippova, "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?," in *Proc. 3rd Blackbox NLP Workshop Analyz. Interpret. Neural Netw. NLP*, 2020 pp. 149–155.

[34] S. Moon, P. Shah, A. Kumar, and R. Subba, "Memory graph networks for explainable memory-grounded question answering," in *Proc. 23rd Conf. Comput. Natural Lang. Learn.*, Nov. 2019, pp. 728–736. [Online]. Available: https://aclanthology.org/K19-1068

[35] Y. Cui, T. Liu, W. Che, Z. Chen, and S. Wang, "Teaching machines to read, answer and explain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1483–1492, 2022.

[36] B. Wu, Z. Zhang, and H. Zhao, "Graph-free multi-hop reading comprehension: A select-to-guide strategy," 2021, arXiv:2107.11823.

[37] Z. Lin, F. Yang, X. Wu, J. Su, and X. Wang, "A feedback-enhanced two-stage framework for judicial machine reading comprehension," *Eng. Appl. Artif. Intell.*, vol. 123, 2023, Art. no. 106178.

[38] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," 2016, *arXiv:1612.08220*.

[39] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber, "Pathologies of neural models make interpretations difficult," in *Proc. 2018 Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3719–3728.

[40] Y. Ju et al., "Enhancing multiple-choice machine reading comprehension by punishing illogical interpretations," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3641–3652.

[41] S. Min et al., "Rethinking the role of demonstrations: What makes in-context learning work?," in *Proc. 2022 Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 11048–11064. [Online]. Available: https://aclanthology.org/2022.emnlp-main.759

[42] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.

[43] B. Wang et al., "Towards understanding chain-of-thought prompting: An empirical study of what matters," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguist.*, 2023, vol. 1, pp. 2717–2739.

[44] P. Lu et al., "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 2507–2521.

[45] A. K. Lampinen et al., "Can language models learn from explanations in context?," 2022, *arXiv:2204.02329*.

[46] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

[47] Z. Tang et al., "Explain-then-translate: An analysis on improving program translation with self-generated explanations," in *Find. Assoc. Comput. Linguist.: EMNLP*, 2023, pp. 1741–1788.

[48] H. Gao et al., "Self-explanation prompting improves dialogue understanding in large language models," 2023, *arXiv:2309.12940*.

[49] S. Li et al., "Explanations from large language models make small reasoners better," 2022, *arXiv:2210.06726*.

[50] E. Zelikman, Y. Wu, J. Mu, and N. Goodman, "STaR: Bootstrapping reasoning with reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 15476–15488.

[51] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju, "Post hoc explanations of language models can improve language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 65468–65483.

[52] M. Turpin, J. Michael, E. Perez, and S. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 74952–74965.

[53] H.-Y. Huang, E. Choi, and W.-T. Yih, "FlowQA: Grasping flow in history for conversational machine comprehension," in *Proc. 7th Int. Conf. Learn. Representations*, 2019.

[54] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[55] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[56] H. Wang et al., "Evidence sentence extraction for machine reading comprehension," in *Proc. 23rd Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2019, pp. 696–707.

[57] K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie, "Dream: A challenge data set and models for dialogue-based reading comprehension," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 217–231, 2019.

[58] K. Sun, D. Yu, D. Yu, and C. Cardie, "Investigating prior knowledge for challenging chinese machine reading comprehension," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 141–155, 2020.

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Representations*, 2019.

[60] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

Hongshen Xu (Graduate Student Member, IEEE) received the B.Eng. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2019. He is currently working toward the Ph.D. degree with X-LANCE Laboratory, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include natural language understanding, alignment, and hallucination of large language modeling.

Lu Chen is currently an Assistant Research Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), Shanghai, China. He has authored or coauthored more than 30 journal articles (e.g., IEEE/ACM Transactions) and peer-reviewed conference papers (e.g., ACL, EMNLP, NAACL, AAAI), one of them was selected as COLING2018 area chair Favorites. His research interests include dialogue systems, question answering, and natural language processing.

Liangtai Sun received the B.Eng. degree in computer science in 2022 from Shanghai Jiao Tong University, Shanghai, China, where he is currently working toward the M.Eng. degree with X-LANCE Laboratory, Department of Computer Science and Engineering. His research interests include dialogue systems, and large language model for science.

Ruisheng Cao received the B.Eng. and M.Eng. degrees in computer science in 2018 and 2021, respectively, from Shanghai Jiao Tong University, Shanghai, China, where he is currently working toward the Ph.D. degree with X-LANCE Laboratory, Department of Computer Science and Engineering. His research interests include semantic parsing, code generation, spoken language understanding, and dialogue systems.

Da Ma received the B.Eng. and M.Eng. degrees in computer science in 2019 and 2022, respectively, from Shanghai Jiao Tong University, Shanghai, China, where he is currently working toward the Ph.D. degree with X-LANCE Laboratory, Department of Computer Science and Engineering. His research interests include dialogue management, and dialogue systems.

Kai Yu (IEEE, Senior Member)received the B.Eng. and M.Sc. degrees from Tsinghua University, Beijing, China, in 1999 and 2002, respectively, and the Ph.D. degree from Cambridge University, Cambridge, U.K., in 2006. He is currently a Professor with Computer Science and Engineering Department, Shanghai Jiao Tong University, Shanghai, China. He joined Machine Intelligence Laboratory, Engineering Department, Cambridge University, U.K. His research interests mainly include the area of speech-based human machine interaction including speech recognition, synthesis, language understanding and dialogue management. He is a member of the IEEE Speech and Language Processing Technical Committee.